VISIBLE AND INFRARED OBJECT TRACKING BASED ON MULTIMODAL HIERARCHICAL RELATIONSHIP MODELING

RUI YAO✉,[1,2], JIAZHU QIU[1,2], YONG ZHOU[1,2], ZHIWEN SHAO[1,2], BING LIU[1,2], JIAQI ZHAO[1,2] AND HANCHENG ZHU[1,2]

[1]The School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China, [2]The Engineering Research Center of Mine Digitization, Ministry of Education of the Peoples Republic of China, Xuzhou, China
e-mail: ruiyao@cumt.edu.cn, ts21170058p31@cumt.edu.cn

(Received January 14, 2024; revised January 26, 2024; accepted February 7, 2024)

ABSTRACT

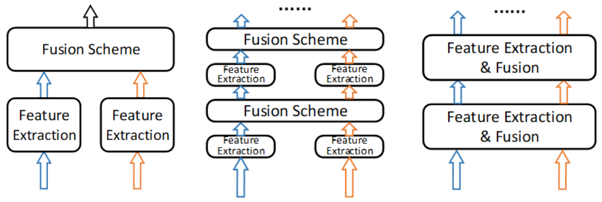
Visible RGB and Thermal infrared (RGBT) object tracking has emerged as a prominent area of focus within the realm of computer vision. Nevertheless, the majority of existing RGBT tracking methods, which predominantly rely on Transformers, primarily emphasize the enhancement of features extracted by convolutional neural networks. Unfortunately, the latent potential of Transformers in representation learning has been inadequately explored. Furthermore, most studies tend to overlook the significance of distinguishing between the importance of each modality in the context of multimodal tasks. In this paper, we address these two critical issues by introducing a novel RGBT tracking framework centered on multimodal hierarchical relationship modeling. Through the incorporation of multiple Transformer encoders and the deployment of self-attention mechanisms, we progressively aggregate and fuse multimodal image features at various stages of image feature learning. Throughout the process of multimodal interaction within the network, we employ a dynamic component feature fusion module at the patch-level to dynamically assess the relevance of visible information within each region of the tracking scene. Our extensive experimentation, conducted on benchmark datasets such as RGBT234, GTOT, and LasHeR, substantiates the commendable performance of our proposed approach in terms of accuracy, success rate, and tracking speed.

Keywords: Feature fusion, Multimodal, RGBT tracking, Transformer.

INTRODUCTION

Visible RGB and Thermal infrared (RGBT) object tracking is an emerging direction in the field of object tracking, aiming to exploit the complementary advantages of visible modality and infrared modality to overcome environmental interference and obtain richer feature representations. There are significant modality differences between visible and infrared images.

Due to visible images and infrared images being captured by different spectral cameras, they undergo significantly different imaging processes and possess distinct wavelength ranges, resulting in notable modality differences. The first critical challenge in RGBT object tracking research is how to overcome this heterogeneity between different modalities. Current RGBT tracking methods often utilize dual-branch networks based on convolutional neural networks and employ a fusion strategy to address this heterogeneity issue. There are roughly three categories of methods for fusing multimodal features: In the first category, as shown in Fig. 1(a), fusion is performed only on high-level features extracted from the two branches using a fusion strategy such as concatenation, element-wise addition, or attention mechanisms Li *et al.* (2019b); Mei *et al.* (2021) . In the second category, as shown in Fig. 1(b), a progressive fusion approach is used to fuse features from multiple layers alternately during feature extraction and feature fusion Xiao *et al.* (2022). However, these methods still have some limitations. Firstly, due to potential imperfect alignments in multimodal images, simple linear operations may lead to the loss of discriminative information in the features. Secondly, the fusion of visible and infrared features often focuses on feature-level fusion at higher layers, lacking early-stage interaction, which may lead to the loss of some important low-level semantic details. Additionally, existing Transformer-based methods typically use separate Transformer encoder and decoder layers for feature enhancement and interaction operations, simply stacking convolutional layers and Transformer layers without fully utilizing its advantages in modeling long-term dependencies. Addressing these issues, we propose a Multimodal Hierarchical Relationship Modeling (MHRM) method, as shown in Fig. 1(c). It utilizes a multi-layer Transformer encoder structure to establish a multi-directional and free information flow, connecting visible light-infrared image pairs and template-search image pairs. During early-stage feature learning in the image, feature interaction and fusion are performed simultaneously. In the interaction process, we model intra-modal fine-grained information relationships, extracting modality-related and discriminative features.



(a) Dual-Steam (b) Progressive (c) Hierarchical

Fig. 1: *Comparison of three fusion methods.*

The second key challenge in RGBT tracking is determining the importance of infrared modality information relative to visible modality information. The complementarity of infrared modality information to visible modality information varies depending on factors like lighting, object shape, size, and occlusion. The interaction between the two modalities is not always effective for visual tasks. Therefore, in practical scenarios, the importance of infrared and visible information differs across videos, frames, and even different regions within the same frame. Thus, discerning whether infrared information enhances visible information and predicting the extent of this enhancement is crucial. However, existing methods often treat both modalities equally and overlook the varying importance of the two modalities for tracking tasks. To address this challenge, we design a patch-level Dynamic Component Fusion Module (DCFM), dynamically solving the importance of visible information for each region in the tracking scene. It adaptively adjusts the interaction between visible and infrared information during tracking to better adapt to complex tracking scenarios. To quantify the importance, we introduced a illumination decoupling network to compute the illumination of visible images and used a trainable neural network to calculate weights for each block. These weights are assigned to describe the importance of visible information within each region. Through patch-level weight allocation, it assigns more reasonable weight proportions to regions containing objects, highlighting the object while reducing the background's impact.

In summary, the main contributions of this paper are as follows:

- We propose a single-stream RGBT tracking framework based on multi-modal hierarchical relationship modeling. By stacking multiple layers of Transformer encoders, we establish a multi-directional and free information flow connecting visible and infrared image pairs, progressively aggregating and fusing multi-modal image features at multiple stages of feature learning.
- We design a patch-level dynamic compoment fusion module based to dynamically solve the importance of visible information for each region in the tracking scene. It adaptively adjusts the interaction between visible and infrared information during tracking to better adapt to complex tracking scenarios.
- Extensive experiments on the RGBT234, GTOT, and LasHeR datasets demonstrate that our method achieves competitive performance in terms of precision, success rate, and tracking speed.

RELATED WORK

RGBT OBJECT TRACKING

In recent years, deep learning-based methods have dominated the field of RGBT object tracking. Gao *et al.* propose a deep adaptive fusion network with multiple fusion modules connected to each layer for fusing visible modal features, infrared modal features, and output features from the upper layer to achieve deep fusion of features Gao *et al.* (2019). Wang *et al.* design a cross-modal pattern propagation network to construct inter-modality propagation relationships through affinity correlation, and to mine and exploit potential mode cues for better feature representation Wang *et al.* (2020). Zhang *et al.* propose a method based on attribute-driven representation to represent and aggregate the features of each class of attribute branches separately to effectively predict the attributes in the tracking process Zhang *et al.* (2021a).

TRANSFORMER-BASED OBJECT TRACKING

Chen *et al.* design a tracking method based on Transformer structure Chen *et al.* (2021), where the decoder replaces the correlation operation in the traditional Siamese network framework. Xiao *et al.* propose an attribute-based progressive fusion network that uses a stacked Transformer encoder-decoder structure, where the encoder performs feature enhancement and the decoder performs feature fusion for different branches of features Xiao *et al.* (2022). Ye *et al.* propose a one-stream tracking framework that unifies feature learning and relational modeling by bridging template-search image pairs with a bi-directional information stream established through the Transformer Ye *et al.* (2022). Zhu *et al.* propose a visual prompt multimodal tracking framework that uses modal complementary cueers to generate effective visual prompts, inputting a single-stream Transformer backbone to eliminate the need to design additional network branches Zhu *et al.* (2023).

FEATURE FUSION

Zhu *et al.* design a trident fusion network for RGBT tracking, which fully exploits multilayer depth features by deploying multimodal information and recursively aggregating features from all convolutional layers using a dense feature aggregation module Zhu *et al.* (2022). Song *et al.* use a cross-attention structure to fuse ultrasound data with MRI image data Song *et al.* (2021). Zhang *et al.* propose a fusion sub-network for semantic segmentation of RGBT tracking, which adaptively obtains the weights of different modalities by bridging first and fusing later strategy with multiple channel weighted summation modules Zhang *et al.* (2021c). Meng *et al.* proposes a human interaction understanding framework that blends local and contextual representations with deep graphical architectures to facilitate the understanding of human-computer interaction Meng *et al.* (2023).

METHOD

NETWORK ARCHITECTURE

The proposed RGBT tracking method consists of three stokenes: Multimodal Hierarchical Relationship Modeling (MHRM), Dynamic Component Fusion Module (DCFM) and prediction head. The specific structure is shown in Fig. 2. Multiple ViT Dosovitskiy *et al.* (2021) encoders are used to form the backbone of the Siamese network, which is used to perform feature learning and interaction between the template and the search image, and between the visible and infrared modalities. In the encoding and weighting addition phase, the DCFM is used to compute the weight of the visible modality for each region of the visible image, in order to distinguish the different levels of importance of the visible and infrared images. Finally, the obtained visible and infrared search region features are fused again and reshaped into spatial features for input to the prediction head for subsequent object classification and regression.

MULTIMODAL HIERARCHICAL RELATIONSHIP MODELING

We design a MHRM module to incrementally aggregate and fuse image features in multiple stages of image feature learning through multiple stacked Transformer encoders. Unlike other Transformer-based methods, we eschew the use of decoder structures as a means of feature interaction and instead use only encoder structures, combining multimodal inputs into one feature sequence that is fed into the encoder structure simultaneously. In the multilayer encoder structure, a free flow of information is constructed by self-attention, and the visible-infrared image pairs are connected by a multi-directional information flow. Multimodal information is directed to each other for feature extraction, and each token embedding in the input sequence can complete the global interaction between two pairs. The proposed MHRM structure is shown in the middle part of Fig. 2. A pair of visible and infrared images of a frame of a video sequence in the RGBT dataset is input, and then the images are cropped to obtain the visible template and search image as well as the infrared template and search image. First, we divide the template image and the search image separately. According to the size, the template image is divided into $n \times n$ patches and the search image is divided into $N \times N$ patches, and then the above image patches are sorted into a sequence of patches to obtain the visible template patch sequence $z^v = [z_1^v; z_2^v; \cdots; z_{n^2}^v]$, the visible search patch sequence $z^i = [z_1^i; z_2^i; \cdots; z_{n^2}^i]$, the infrared template patch sequence $x^v = [x_1^v; x_2^v; \cdots; x_{N^2}^v]$ and the infrared search patch sequence $x^i = [x_1^i; x_2^i; \cdots; x_{N^2}^i]$. The linear projection layer flattens z^v, z^i, x^v and x^i to 2-dimensional features, while adding the learnable position embedding p_z and p_x, to mark the position information of each patch. The projection layer outputs the token embedding sequences Z and X, in the case of visible modality, the process can be described as:

$$Z^v = [z_1^v P; z_2^v P; \cdots; z_{n^2}^v P] + p_z, \tag{1}$$

$$X^v = [x_1^v P; x_2^v P; \cdots; x_{N^2}^v P] + p_x. \tag{2}$$

where P is the learnable parameter of the linear projection layer. The same can be done for the infrared token embedding sequence. According to the distribution of token embeddings, the visible and infrared token embeddings are sequentially cross-arranged and concatenated into a sequence, which is then fed in parallel into a MHRM consisting of L Transformer encoders. The encoder structure uses the ViT structure that has been applied many times to downstream tasks, with some modifications to make it more suitable for multimodal tasks, as

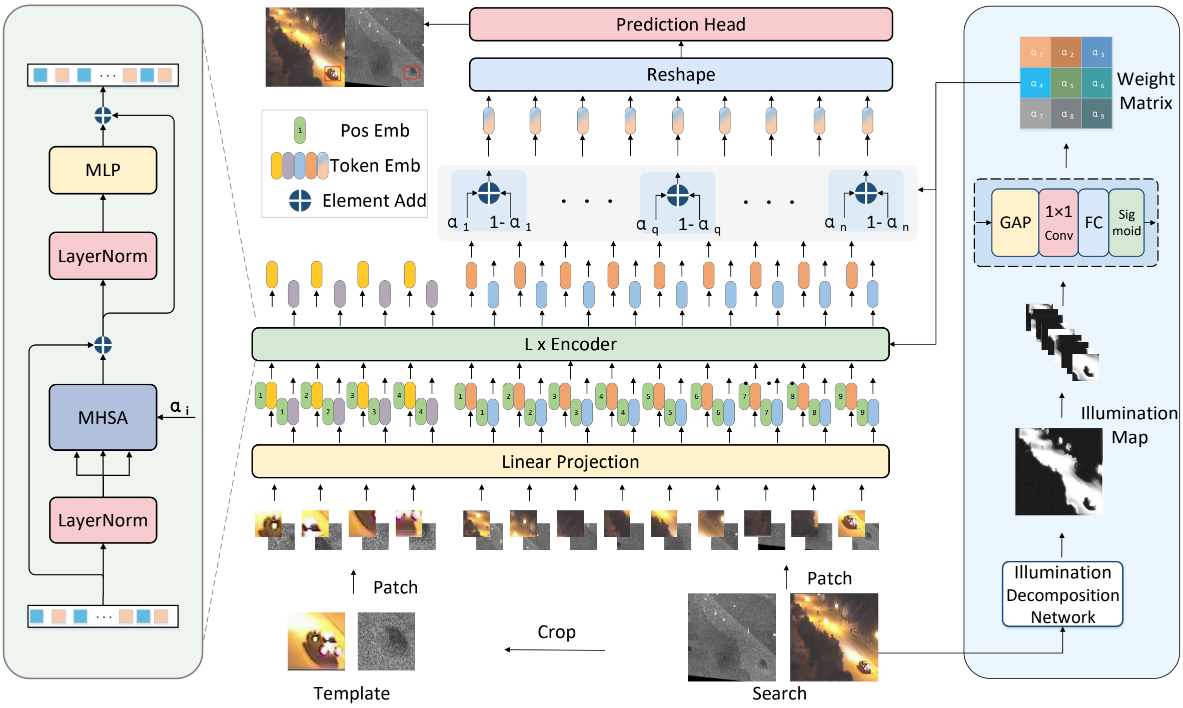


Fig. 2: *The framework of multimodal hierarchical relationship modeling tracking model. It includes three parts: multimodal hierarchical relationship modeling, dynamic component feature fusion and prediction head.*

shown in the left part of Fig. 2. The encoder consists of two layer normalization followed by a multi-head self-attention and a multi-layer perceptron layer, during which two residual connections are made. ViT provides a variety of publicly available pre-training models, which can greatly improve the efficiency of our training models. In the encoder, the input token embeddings sequence consisting of two modalities is subjected to multiple self-attention operations. Unlike the cross-attention of two inputs, the self-attention is a process of interaction between two input token embeddings, which enhances their own features by generating an attention matrix. In this process, not only enhancing of the two modalities' own feature representations, but also the interactive fusion between the template and the search, and between the visible and infrared images are carried out. Furthermore, the weight allocation network in Section 3.3 also plays a role in this process. After obtaining the weight matrix W, we calculate the ratio matrix between W and $(1-W)$ before encoding, and then multiply it with the infrared token sequence. Also, since this phase is performed simultaneously, the fusion of visible and infrared information is discriminative and instructive in the training process. Compared to cross-attention, self-attention using cascaded features makes the whole framework highly parallelized. Although the input to ViT is still a visible-infrared image pair, the inference speed is less affected by the highly parallel structure.

We use a token subsequence $[Z_j^v;Z_j^i;X_j^v;X_j^i]$ to illustrate the principle of the method. This process can be analyzed from two perspectives. First, the formula for the attention mechanism can be expressed as:

$$A = Softmax(\frac{[Q_z^v;Q_z^i;Q_x^v;Q_x^i][K_z^v;K_z^i;K_x^v;K_x^i]^\top}{\sqrt{d_k}})\cdot[V_z^v;V_z^i;V_x^v;V_x^i], \tag{3}$$

From the perspective of multimodal relationship, the attention weight map calculation process can be expressed as follows:

$$\begin{aligned} W &= [Q_z^v;Q_z^i;Q_x^v;Q_x^i][K_z^v;K_z^i;K_x^v;K_x^i]^\top \\ &= [Q_{\{z,x\}}^v {K_{\{z,x\}}^v}^\top;Q_{\{z,x\}}^v {K_{\{z,x\}}^i}^\top;Q_{\{z,x\}}^i {K_{\{z,x\}}^v}^\top;Q_{\{z,x\}}^i {K_{\{z,x\}}^i}^\top] \\ &\triangleq [W_{vv},W_{vi};W_{iv},W_{ii}]. \end{aligned} \tag{4}$$

where W_{iv} can be considered as a measure of the similarity between visible and infrared images, resulting in the self-attention output:

$$\begin{aligned} A &= [W_{vv},W_{vi};W_{iv},W_{ii}]\cdot[V_{\{z,x\}}^v;V_{\{z,x\}}^i] \\ &= [W_{vv}V_{\{z,x\}}^v + W_{vi}V_{\{z,x\}}^i;W_{iv}V_{\{z,x\}}^v + W_{ii}V_{\{z,x\}}^i], \end{aligned} \tag{5}$$

where $W_{iv}V_{\{z,x\}}^v$ is responsible for the fusion between visible and infrared modalities, while $W_{vv}V_{\{z,x\}}^v$ and $W_{ii}V_{\{z,x\}}^i$ are feature aggregation operations for image itself. Therefore, the global relationship modeling of the L-layer Transformer encoders achieves a more adequate perceptual fusion of visible and infrared information.

Similarly, from the perspective of relationship between template and the search image, the attention weight map calculation process can be expressed as follows:

$$\begin{aligned} W &= [Q_z^{\{v,i\}}K_z^{\{v,i\}\top}; Q_z^{\{v,i\}}K_x^{\{v,i\}\top}; Q_x^{\{v,i\}}K_z^{\{v,i\}\top}; Q_x^{\{v,i\}}K_x^{\{v,i\}\top}] \\ &\triangleq [W_{zz}; W_{zx}; W_{xz}; W_{xx}], \end{aligned} \tag{6}$$

where W_{zx} can be considered as a measure of the similarity between template and search images, resulting in the self-attention output:

$$\begin{aligned} A &= [W_{zz}, W_{zx}; W_{xz}, W_{xx}] \cdot [V_z^{\{v,i\}}; V_x^{\{v,i\}}] \\ &= [W_{zz}V_z^{\{v,i\}} + W_{zx}V_x^{\{v,i\}}; W_{xz}V_z^{\{v,i\}} + W_{xx}V_x^{\{v,i\}}], \end{aligned} \tag{7}$$

where $W_{xz}V_z^{\{v,i\}}$ is responsible for the relational interaction between the template image and the search image, $W_{xx}V_x^{\{v,i\}}$ is equivalent to feature aggregation by attention of the image itself.

DYNAMIC COMPONENT FEATURE FUSION

We focus on how to assign reasonable weights to the multimodal information to regulate the importance of visible and infrared information in the whole tracking task, and thus guide the interaction between the different modalities. We design a patch-level DCFM by introducing an illumination decomposition network to obtain the input visible image illumination map, and then dynamically derive the corresponding weights for each patch by a trainable neural network.

Since there may be significant scene changes between videos and even between frames, we adjust the weighting of the interactions between the different modes to a dynamic value. Since there may be significant illumination differences in different regions of the same frame, we design a Dynamic Component Fusion Module (DCFM) at image patch-level by introducing an illumination decoupling network to obtain the input visible image illumination map, and then dynamically derive the corresponding weights for each patch by a trainable neural network.

DCFM estimates a deterministic value $\alpha \in (0,1)$ to describe the trustworthiness of the visible information in each region by measuring the illumination information of the visible image. α and $(1-\alpha)$ will be used as modality weights to dynamically guide the interaction of visible and infrared modalities throughout. Specifically, we directly refer to the illumination decoupling network in the publicly available pre-trained weight model KinD++ Zhang *et al.* (2021d). Based on the Retinex illumination enhancement theory, we set two branches for the visible image to decompose the illumination component I and the reflectance component R of the visible image, respectively, denoted as:$S = R \cdot I$. We keep only the illumination component as illumination map I for subsequent operations. For the actual tracking task, we only need a deterministic value to regulate the multimodal fusion. Therefore, similar to TNet Cong *et al.* (2022), we set a trainable network to map the illumination map to (0,1) to describe the trustworthiness of the visible illumination. Specifically, the illumination map is divided into regions according to the image patch partitioning rule, then resized by global average pooling, 1×1 convolutional transformation of the channels, and then, finally, the features are mapped to a specific fractional value by a fully connected layer and a sigmoid activation function. The process can be expressed as follows:

$$\alpha_i = \sigma(FC(Conv(GAP(I_i)))), \tag{8}$$

where I_i denotes the feature of the i-th region, $Conv$ and FC represents the convolution and fully connected layer, GAP and σ denotes the global average pooling and sigmoid activation function. α denotes the final weight.

After calculating α for each region, we can obtain the weight matrix W for the search image. W initializes the weights for both visible and infrared modalities across the entire network and adjusts them through a trainable network. The regions of operation for W are the encoder within the main backbone and the final weighted summation module. It globally regulates the fusion-related effects for multimodal integration.

The image patch-level weight assignment can effectively regulate the global multimodal information fusion. To a certain extent, assigning a more reasonable weight to the region patches containing the objects can achieve the purpose of highlighting the object and weakening the background. The is input to the encoder structure in the backbone and the final weighted addition module to globally regulate the fusion-related effects of multimodal.

PREDICTION HEAD AND LOSS FUNCTION

A sequence of token embeddings containing multimodal information is reshaped into a spatial feature map, which are then fed into a fully convolutional network consisting of m convolutional-normalized-ReLU activation function layers. The response maps M and local offsets O are output to obtain the final predicted classification results and object coordinates. In the training process, the entire tracking network architecture uses both classification and regression loss functions to achieve the best training results. We use weighted focal loss Law and Deng (2018) as the classification loss, GIoU loss Rezatofighi *et al.* (2019) as the regression loss, and mean absolute error loss.

Weighted focal loss adjusts the training focus onto challenging samples by dynamically modifying the weights of easily distinguishable samples during the training process. Its calculation formula can be described as follows:

$$\mathcal{L}_{cls} = -\sum_{x}^{H}\sum_{y}^{W}\begin{cases} (1-M_{xy})^{\beta}\log(M_{xy}), if \hat{M}_{xy}=1 \\ (1-\hat{M}_{xy})^{\eta}(M_{xy})^{\beta}\log(1-M_{xy}), otherwise \end{cases}, \tag{9}$$

where M_{xy} is the prediction score at position (x,y) in the predicted response map, $\hat{M}_{xy}$ denotes the truth heat map generated using Gaussian kernel, β and η is the hyperparameter set to 2 and 4 respectively during training. The IoU loss employs the Intersection over Union (IoU) metric to address the issue where bounding boxes with the same L-distance between predicted and ground truth boxes have different IoU values, making it challenging to optimize using IoU alone. However, when two bounding boxes do not intersect (i.e., IoU=0), the loss value is 0, which prevents gradient backpropagation. Hence, the Generalized IoU (GIoU) loss is used. It introduces a minimum enclosing box to confine the overlap range. The calculation formula is as follows:

$$\mathcal{L}_{GIoU} = IoU + \frac{A^{c}-\mu}{A^{c}}, \tag{10}$$

where A^{c} represents the minimum enclosing box area of the real target bounding box and the predicted bounding box, while μ stands for the union area of the real bounding box and the predicted bounding box. The total loss of the network is described as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{1}\mathcal{L}_{1} + \lambda_{2}\mathcal{L}_{GIoU}, \tag{11}$$

where λ_1 and λ_2 are the equilibrium parameters, which are set to 2 and 5 in the experiment.

EXPERIMENTS

DATASETS AND METRICS

GTOT includes 50 pairs of highly aligned visible and infrared videos, which were captured in different scenes and conditions. Each frame contains manually annotated data, including the coordinates of the object's bounding box and attributes indicating challenging conditions Li *et al.* (2016).

RGBT234 is an extension of the RGBT210 dataset, comprising 234 pairs of highly aligned visible and infrared videos. It also includes manually annotated object bounding boxes and attributes indicating challenging conditions. The annotations are more accurate, and the attributes are richer, taking into account various environmental challenges Li *et al.* (2019a).

LasHeR is a large RGBT dataset that consists of 1224 pairs of visible and infrared videos, featuring greater scene complexity. Among these, 979 video sequences are allocated to the training set, while 245 sequences are allocated to the test set Li *et al.* (2021).

In the evaluation of the GTOT, RGBT234, and LasHeR test sets, we use the same two evaluation metrics: Precision Rate (PR) and Success Rate (SR). We calculate the center position error between predicted bounding boxes and ground truth bounding boxes for all frames and set a threshold, where the CLE is defined as:

$$\rho = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}, \tag{12}$$

where (x_1,y_1) and (x_2,y_2) indicates the center point coordinates of the predicted bounding box and the ground truth. PR represents the percentage of all frames whose center position error is less than this threshold. Similarly, we compute the overlap between the predicted bounding box and the groundtruth for all frames and set a threshold, where the overlap is defined as:

$$O(a,b) = \frac{|a\cap b|}{|a\cup b|}, \tag{13}$$

where a and b indicates the predicted bounding box and ground truth. SR is the percentage of all frames whose overlap is greater than the threshold, i.e. the percentage of frames successfully tracked.

EXPERIMENT SETTING

We conduct experiments on a computer equipped with an NVIDIA GTX 4090 GPU, running the Ubuntu 20.04 operating system.

During the training process, the OSTrack-384 is used as a baseline. Firstly, the pretrained model parameters based on the ViT with MAE He *et al.* (2022) are loaded, and the number of encoders, denoted as L, is set to 12. The encoder model parameters are initialized. Subsequently, the DCFM parameters are initialized based on the Retinex-based illumination decoupling network. The entire model is trained using the LasHeR dataset train sets, and data augmentation strategies are employed during training, including operations like flipping, rotation, brightness jitter, and so on. The initial learning rate for the encoder backbone is set to 4×10^{-5}, while the learning rate for other network structures in the model is set to 4×10^{-4}. The Adam W optimizer is used to optimize the model with a weight decay of 10^{-4}, and the training iteration count is set to 100.

During testing, the pretrained model parameters are loaded, and model parameters are fixed. The classification map is simply multiplied by a Hann window of the same size. The bounding box with the highest score after multiplication is selected as the tracking result.

EVALUATION ON RGBT TRACKING DATASETS

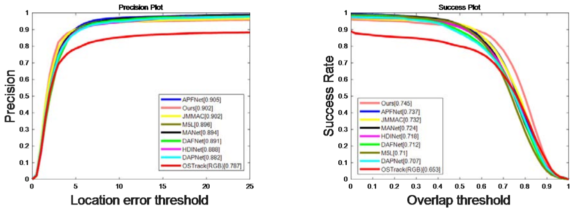


Fig. 3: *Evaluation results on the GTOT dataset.*

GTOT Dataset. We compare the our method (Ours) with several state-of-the-art tracking methods in terms of both accuracy and success rate, including the RGBT tracking methods APFNet Xiao *et al.* (2022), MANet Li *et al.* (2019b), JMMAC Zhang *et al.* (2021b), M5L Tu *et al.* (2022), HDINet Mei *et al.* (2021), DAFNet Gao *et al.* (2019), DAPNet Zhu *et al.* (2019) and the traditional tracking method OTrack Ye *et al.* (2022). The comparison of tracking results on the GTOT dataset is shown in Fig. 3. The comparison results show that our method shows superior accuracy and success rate compared with most of the tracking methods, and achieves the best performance in the success rate metric. It achieves 74.5%. The accuracy achieves 90.2%. The accuracy and success rate are 11.5% and 9.2% higher than the baseline method OTrack, respectively. The success rate is 0.8% higher and the accuracy is only 0.3% lower than that of the state-of-the-art method APFNet. In addition, the proposed method is based on SiameseFC, which has a significant speed advantokene over APFNet based on MDNet Nam and Han (2016), and the speed comparison results are shown in detail in Section 4.5.

RGBT234 Dataset. The comparison of tracking results on the RGBT234 dataset is shown in Fig. 4. The comparative results show that our method also achieves better results on RGBT234. The best performance is achieved in the success rate index, which reaches 59.9%. The accuracy achieves 80.5%. The accuracy and success rate are 5.6% and 3.6% higher than the baseline method OSTrack, respectively. The success rate is 2% higher and the accuracy is 2.2% lower than that of the state-of-the-art method APFNet.

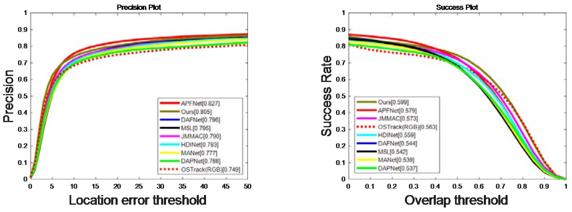


Fig. 4: *Evaluation results on the RGBT234 dataset.*

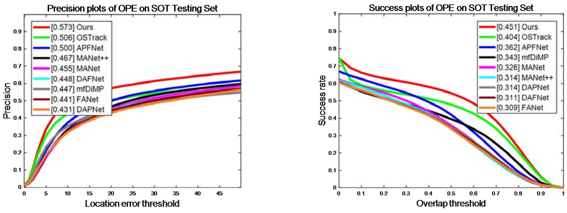


Fig. 5: *Evaluation results on the LasHeR dataset.*

LasHeR Dataset. We compare our method (Ours) with several RGBT tracking methods, including APFNet Xiao *et al.* (2022), MANet Li *et al.* (2019b), MANet++ Lu *et al.* (2021), DAFNet Gao *et al.* (2019), DAPNet Zhu *et al.* (2019), FANet Zhu *et al.* (2021), mfDiMP Zhang *et al.* (2019) and OSTrack Ye *et al.* (2022). The comparison results are shown in Fig. 5. The results show that our method achieves best performance in both accuracy and success rate. Compared with the method APFNet, the accuracy is improved by 7.3% and the success rate is increased by 8.9%. According to the analysis, the LasHeR dataset has more complex scenes and challenging environments, and the proposed method focuses more on modulating the enhancement of visible modality by infrared information, so it is more advantokeneous when facing more challenging environments.

Visualization. The tracking visualization results of GTOT and RGBT234 subsequences are shown in Fig. 6, and the visualization of the LasHeR subsequencesis shown in Fig. 7. The results show that our method compared to APFNet and OSTrack, the tracking bounding box results are closer to GroundTruth and less prone to tracking drift when dealing with video sequences in poor environments such as nighttime or low-light.

EVALUATION OF ATTRIBUTE CHALLENGE

GTOT dataset primarily consists of seven attributes, corresponding to seven external environmental challenges. These challenges include target occlusion (OCC), scale variation (LSV), fast

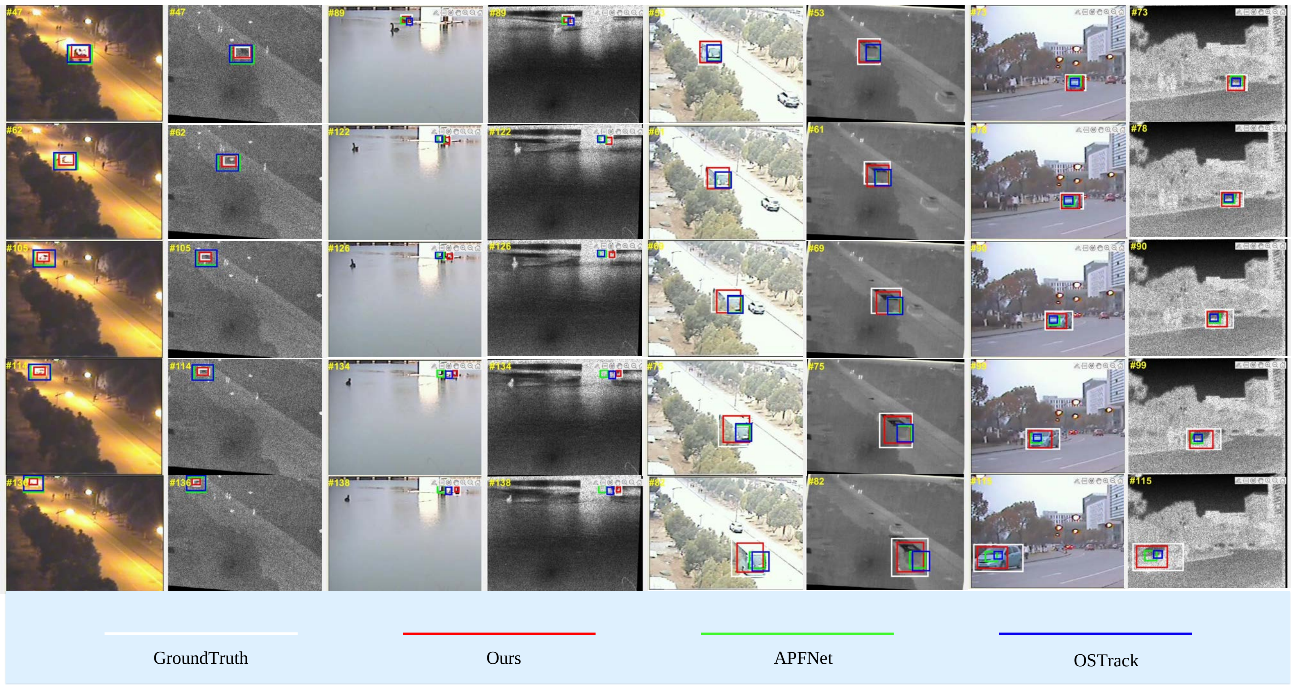


Fig. 6: *Visualizations of four video sequence on GTOT and RGBT234 datasets.*

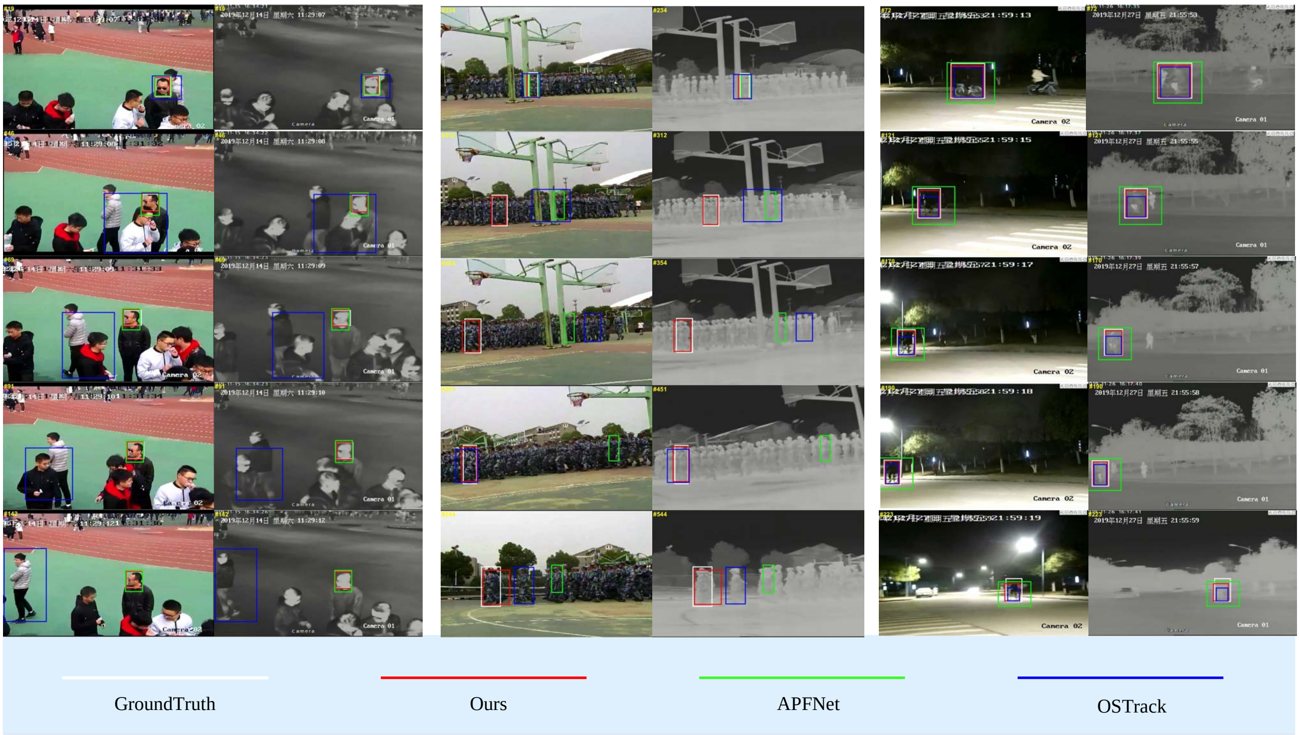


Fig. 7: *Visualizations of three video sequence on LasHeR datasets.*

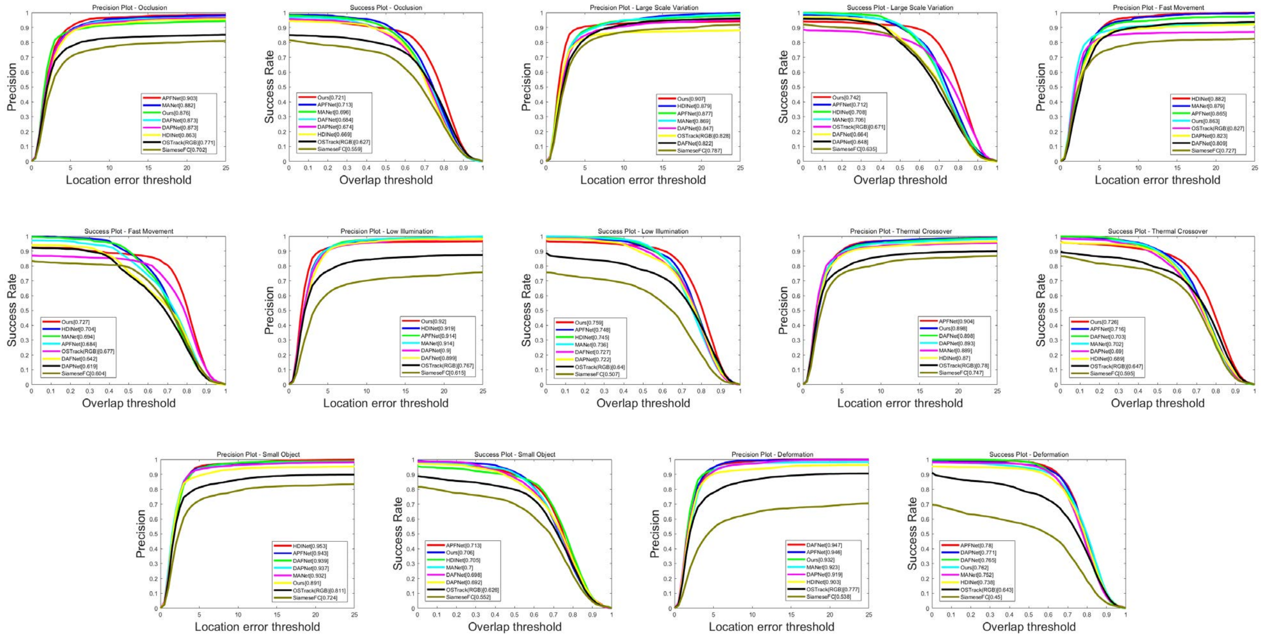


Fig. 8: *Attribute-based PR/SR scores on GTOT dataset.*

motion (FM), low illumination (LI), thermal crossover (TC), small objects (SO), and object deformation (DEF). To address these attribute challenges, we conducted precision and success rate comparisons of seven tracking methods on the GTOT dataset, as shown in Fig. 8. The figure shows that our method achieves the best results for both SR and PR in the LSV and LI attributes, and SR alone achieves the best results in the OCC, FM, and TC attributes, and PR is mostly in the second position. There are also cases of PR or SR in second place in SO and DEF. In summary, our method maintains the best or second-best performance on most attributes' two metrics, especially in scale variation, low illumination. Compared to attribute-based method APFNet, our method also shows certain advantages. Experimental results confirm that our method excels in dealing with external environmental challenges.

ABLATION STUDY

Module Ablation Experiments. In order to verify the effectiveness of proposed MHRM and DCFM, the entire network was disassembled and combined. To further validate the effectiveness of DCFM, Static Component Fusion Module (SCFM) was set and α was manually set to 0.6. Ablation experiments were conducted on two datasets, RGBT234 and GTOT. The comparison results are shown in Table 1. The results show that the tracking model with the addition of the MHRM and DCFM can improve the accuracy and success rate, and the weight assignment provided by the DCFM can also improve the performance compared to the static assignment, and the synergy between the two can achieve the best performance improvement. This demonstrates the effectiveness of each of the proposed modules.

Candidate Elimination Ablation Experiments. The Candidate Elimination (CE) module is a key module in the benchmark method OSTrack. We compared the method with the Candidate Elimination (CE) retained to the method with CE removed. The comparison results are shown in Table 1. The experimental results show that the model with CE in place has a slight decrease in both metrics for both datasets. Therefore, we removed CE module. The experimental results indicate that incorporating the CE module leads to a slight decrease in both metrics for both datasets. As a result, we decided to remove the CE module.

EFFICIENCY ANALYSIS

The efficiency analysis experiments were conducted in the same environment. We compare the tracking speed of our method with several tracking methods with better performance, APFNet Xiao *et al.* (2022), MANet Li *et al.* (2019b), and MANet++ Lu *et al.* (2021). The comparison results are shown in Table 2. The results show that the tracking speed of our method can reach 87 FPS, which is sufficient to achieve real-time tracking. Compared with the best method APFNet, the average FPS of our method is 78.9 higher than APFNet. Compared with the fastest

Table 1: *PR/SR scores of ablation experiments on GTOT and RGBT234.*

RGB	T	MHRM	SCFM	DCFM	CE Ye *et al.* (2022)	GTOT	RGBT234
✓						78.7/65.3	74.9/56.3
	✓					64.7/55.1	70.4/51.3
✓	✓	✓				83.6/68.4	77.0/57.5
✓	✓	✓	✓			89.2/73.8	77.2/57.9
✓	✓	✓		✓	✓	90.0/74.2	79.2/59.1
✓	✓	✓		✓		**90.2/74.5**	**80.5/59.9**

method MANet++, the average FPS is improved by 35.6. The data in the table are sufficient to prove that our method performs well in terms of accuracy, success rate and tracking speed.

Table 2: *Comparison of efficiency and real-time performance (PR/SR/FPS) of four methods.*

Method	GTOT	RGBT234
MANet	89.4/72.4/6.2	77.7/53.9/5.9
MANet++	90.1/72.3/52.9	80.0/55.4/50.4
APFNet	90.5/73.7/8.5	82.7/57.9/8.2
Ours	**90.2/74.5/87.5**	**80.5/59.9/87.0**

CONCLUSION

In this paper, we propose an RGBT tracking framework, which uses a stacked Transformer encoders to progressively aggregate and fuse multimodal image features. During the entire multimodal interaction process of the network, a DCFM is used to dynamically solve for the importance of visible information in each region of the tracking scene, thereby regulating the interaction between visible and infrared information in the tracking process. Experimental results on three datasets demonstrate the competitive performance of the proposed method.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62172417, 62272461, 62106268, 62276266 and 62101555, and in part by the Xuzhou Key Research and Development Program under Grant KC22287.

REFERENCES

Chen X, Yan B, Zhu J, Wang D, Yang X, Lu H (2021). Transformer tracking. In: CVPR.

Cong R, Zhang K, Zhang C, Zheng F, Zhao Y, Huang Q, Kwong S (2022). Does thermal really always matter for rgb-t salient object detection? IEEE Transactions on Multimedia .

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR .

Gao Y, Li C, Zhu Y, Tang J, He T, Wang F (2019). Deep adaptive fusion network for high performance rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).

He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022). Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Law H, Deng J (2018). Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV).

Li C, Cheng H, Hu S, Liu X, Tang J, Lin L (2016). Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing 25:5743–56.

Li C, Liang X, Lu Y, Zhao N, Tang J (2019a). Rgb-t object tracking: Benchmark and baseline. Pattern Recognition 96:106977.

Li C, Xue W, Jia Y, Qu Z, Luo B, Tang J, Sun D (2021). Lasher: A large-scale high-diversity benchmark for rgbt tracking. IEEE Transactions on Image Processing 31:392–404.

Li CL, Lu A, Zheng AH, Tu Z, Tang J (2019b). Multi-adapter rgbt tracking. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).

Lu A, Li C, Yan Y, Tang J, Luo B (2021). Rgbt tracking via multi-adapter network with hierarchical divergence loss. IEEE Transactions on Image Processing 30:5613–25.

Mei J, Zhou D, Cao J, Nie R, Guo Y (2021). Hdinet:

Hierarchical dual-sensor interaction network for rgbt tracking. IEEE Sensors Journal 21:16915–26.

Meng J, Wang Z, Ying K, Zhang J, Guo D, Zhang Z, Shi JQ, Chen S (2023). Human interaction understanding with consistency-aware learning. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Nam H, Han B (2016). Learning multi-domain convolutional neural networks for visual tracking. IEEE .

Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Song X, Guo H, Xu X, Chao H, Xu S, Turkbey B, Wood BJ, Wang G, Yan P (2021). Cross-modal attention for mri and ultrasound volume registration. Berlin, Heidelberg: Springer-Verlag.

Tu Z, Lin C, Zhao W, Li C, Tang J (2022). M5l: Multi-modal multi-margin metric learning for rgbt tracking. IEEE Transactions on Image Processing 31:85–98.

Wang C, Xu C, Cui Z, Zhou L, Zhang T, Zhang X, Yang J (2020). Cross-modal pattern-propagation for rgb-t tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Xiao Y, Yang M, Li C, Liu L, Tang J (2022). Attribute-based progressive fusion network for rgbt tracking. AAAI .

Ye B, Chang H, Ma B, Shan S, Chen X (2022). Joint feature learning and relation modeling for tracking: A one-stream framework. In: European Conference on Computer Vision. Springer.

Zhang L, Danelljan M, Gonzalez-Garcia A, Van De Weijer J, Shahbaz Khan F (2019). Multi-modal fusion for end-to-end rgb-t tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Zhang P, Wang D, Lu H, Yang X (2021a). Learning adaptive attribute-driven representation for real-time rgb-t tracking 129:2714–2729.

Zhang P, Zhao J, Bo C, Wang D, Lu H, Yang X (2021b). Jointly modeling motion and appearance cues for robust rgb-t tracking 30:3335 – 3347.

Zhang Q, Zhao S, Luo Y, Zhang D, Huang N, Han J (2021c). Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang Y, Guo X, Ma J, Liu W, Zhang J (2021d). Beyond brightening low-light images. International Journal of Computer Vision 129:1013–37.

Zhu J, Lai S, Chen X, Wang D, Lu H (2023). Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Zhu Y, Li C, Luo B, Tang J, Wang X (2019). Dense feature aggregation and pruning for rgbt tracking. In: Proceedings of the 27th ACM International Conference on Multimedia, MM ’19. Association for Computing Machinery.

Zhu Y, Li C, Tang J, Luo B (2021). Quality-aware feature aggregation network for robust rgbt tracking. IEEE Transactions on Intelligent Vehicles 6:121–30.

Zhu Y, Li C, Tang J, Luo B, Wang L (2022). Rgbt tracking by trident fusion network. IEEE Transactions on Circuits and Systems for Video Technology 32:579–92.